

Inverse Ising inference using all the data

Erik Aurell*

ACCESS Linnaeus Centre, KTH, Stockholm, Sweden
and Dept. Computational Biology, AlbaNova University Centre, 106 91 Stockholm, Sweden

Magnus Ekeberg†

Engineering Physics Program
KTH Royal Institute of Technology,
100 77 Stockholm, Sweden

We show that a method based on logistic regression, using all the data, solves the inverse Ising problem far better than mean-field calculations relying only on sample pairwise correlation functions, while still computationally feasible for hundreds of nodes. The largest improvement in reconstruction occurs for strong interactions. Using two examples, a diluted Sherrington-Kirkpatrick model and a two-dimensional lattice, we also show that interaction topologies can be recovered from few samples with good accuracy and that the use of l_1 -regularization is beneficial in this process, pushing inference abilities further into low-temperature regimes.

Introduction: When analyzing systems of interacting elements, distinguishing direct correlations (caused by actual interactions between elements) from indirect correlations (induced through chains of interactions via other elements) is an intrinsically complex task. Versions of this problem come about naturally in biology, sociology, neuroscience and many other fields, and are bound to become more and more important as the amount and diversity of data on large systems will continue to grow. In the Ising model, which has served as a basic starting point for studying such situations in applications [1–3], a set of binary variables $\sigma = \{\sigma_1, \dots, \sigma_N\}$, $\sigma_i = \pm 1$, have the distribution

$$P(\sigma) = \frac{1}{Z} \exp \left(\beta \sum_i h_i \sigma_i + \beta \sum_{i < j} J_{ij} \sigma_i \sigma_j \right) \quad (1)$$

where Z is the partition function, $\beta = 1/T$ the inverse temperature, h_i are the external fields and J_{ij} the pairwise couplings. Given magnetizations $m_i = \langle \sigma_i \rangle$ and pairwise correlations $c_{ij} = \langle \sigma_i \sigma_j \rangle - m_i m_j$ the probability distribution which maximizes the entropy has the Ising model form. The standard *inverse Ising problem* means to compute (approximately, efficiently, or according to other criteria) the parameters h_i and J_{ij} from observed m_i and c_{ij} . The practical interest in inverse Ising, in the context of the present and future data-rich world, is to use it as an information extraction tool alternative and/or superior to measuring correlations. Extending the number of states from two to twenty, spectacular success has been achieved in inferring directly interacting residues (amino acids) in two-component signaling pathways in bacteria [4, 5], and use has also been reported for

protein structure prediction [6, 7]. In this Letter we address the following two questions: (i) can one do better by keeping all the data for reconstruction and not only empirical pairwise correlation functions, and (ii) can such a method be implemented in a computationally efficient manner? The answer is positive on both accounts, using a method inspired by the regularized logistic regression process of Wainwright, Ravikumar and Lafferty [8]. We show in particular that keeping all the data greatly improves reconstruction of an Ising model in the important parameter region of strong interactions.

Maximum log-likelihood, exponential families and computability: We will from now on assume that we have B independent observations $\{\sigma^{(k)}\}_{k=1}^B$ all drawn from (1). The log-likelihood function, given these observations, is

$$l(\{h_i\}, \{J_{ij}\}; \{\sigma^{(k)}\}_{k=1}^B) = \beta \sum_i h_i m_i^{(B)} + \beta \sum_{i < j} J_{ij} (m_i^{(B)} m_j^{(B)} + c_{ij}^{(B)}) - \log Z \quad (2)$$

where $m_i^{(B)}$ and $c_{ij}^{(B)}$ are the empirical first and second moments from B samples. A classical result in statistics states that for exponential families of parameter distributions, of which the Ising model (1) is an instance, the averages of the functions multiplying the model parameters are sufficient statistics [9–11]. This means that “no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter” [12], or, in the case at hand, that inference of the biases h_i and the interaction strengths J_{ij} cannot be done better using all the B samples (NB data points), than by observing just $m_i^{(B)}$ and $c_{ij}^{(B)}$ ($\frac{N(N+1)}{2}$ data points). The optimal estimates (in a maximum likelihood sense) are then given by $\partial_{h_i} \log Z = \beta m_i^{(B)}$ and $\partial_{J_{ij}} \log Z = \beta [m_i^{(B)} m_j^{(B)} + c_{ij}^{(B)}]$. The problem with this solution is that it is unfeasible to compute the partition function exactly in large systems. A whole series of approximations, reviewed in [13], have

* Also at Aalto University School of Science, Helsinki, Finland ;
eaurell@kth.se

† ekeb@kth.se

therefore been developed by expanding in high temperature (small interactions), large external fields or other parameters *cf.* (naive) mean-field (nMF) [14], TAP [14], loop summation [15] and have been further extended using the fluctuation-dissipation theorem [16–18]. It is well-established that all these approximate methods are not accurate when the number of samples is small, nor when the interactions are strong (temperature is low). However, a recent method based on expansion of the system into "clusters" (who's contributions to the estimates of $\{\mathbf{h}, \mathbf{J}\}$ are included or discarded depending on their entropy share) manages to select correctly the parameters from few samples in various low-temperature settings [19], questioning these limitations.

Pseudo-likelihood maximization (without regularization): The conditional probability of one variable σ_r given all the others $\boldsymbol{\sigma}_{\setminus r} = (\sigma_1, \dots, \sigma_{r-1}, \sigma_{r+1}, \dots, \sigma_N)$ is

$$P_{\{\mathbf{h}, \mathbf{J}\}}(\sigma_r | \boldsymbol{\sigma}_{\setminus r}) = \frac{1}{1 + \exp(-2\beta\sigma_r[h_r + \sum_{i \neq r} J_{ir}\sigma_i])} . \quad (3)$$

If σ_r by itself is considered a dependent variable, and the complementary set $\boldsymbol{\sigma}_{\setminus r}$ is taken as independent variables, then the maximum likelihood estimates of the parameters h_r and $\mathbf{J}_r = \{J_{ir}\}_{i \neq r}$, given B samples, minimize

$$f_r(h'_r, \mathbf{J}'_r) = -\frac{1}{B} \sum_{k=1}^B \ln P_{\{\mathbf{h}', \mathbf{J}'\}}(\sigma_r^{(k)} | \boldsymbol{\sigma}_{\setminus r}^{(k)}) . \quad (4)$$

Minimizing these functions f_r for all r simultaneously is not the same as maximizing the total log-likelihood (2). For example, this procedure, which we call pseudo-likelihood maximization, would typically give different estimates $J_{ij}^{*,i}$ and $J_{ij}^{*,j}$ depending on if σ_i or σ_j is considered the dependent variable. We will for definiteness always take the pseudo-likelihood maximization estimate of the coupling constant to be the average $J_{ij}^* = \frac{1}{2} (J_{ij}^{*,i} + J_{ij}^{*,j})$. When the number of samples is large we can substitute sample average with ensemble average, and write

$$\begin{aligned} f_r(h'_r, \mathbf{J}'_r) &\approx \langle -\ln(P_{\{\mathbf{h}', \mathbf{J}'\}}(\sigma_r | \boldsymbol{\sigma}_{\setminus r})) \rangle \\ &= \sum_{\boldsymbol{\sigma}} \ln \left(1 + e^{-2\beta\sigma_r[h'_r + \sum_{i \neq r} J'_{ir}\sigma_i]} \right) P_{\{\mathbf{h}, \mathbf{J}\}}(\boldsymbol{\sigma}) , \end{aligned} \quad (5)$$

with equality expected in the limit. Necessary maximum likelihood conditions (for one of the conditional probabilities) are then

$$\frac{\partial f_r}{\partial J'_{sr}}(h'_r, \mathbf{J}'_r) = \sum_{\boldsymbol{\sigma}} \frac{-2\beta\sigma_s\sigma_r}{e^{2\beta\sigma_r[h'_r + \sum_{i \neq r} J'_{ir}\sigma_i]} + 1} P_{\{\mathbf{h}, \mathbf{J}\}}(\boldsymbol{\sigma}) = 0 \quad (6)$$

and similarly for the variation with respect to an external

field. At the true parameters these equations hold, since

$$\frac{\partial f_r}{\partial J'_{sr}}(h_r, \mathbf{J}_r) = \frac{-\beta}{Z\{\mathbf{h}, \mathbf{J}\}} \sum_{\boldsymbol{\sigma}} \sigma_s \sigma_r \frac{e^{\beta \sum_{i \neq r} h_i \sigma_i + \beta \sum_{i < j, i, j \neq r} J_{ij} \sigma_i \sigma_j}}{\cosh(\beta \sigma_r [h_r + \sum_{i \neq r} J_{ir} \sigma_i])} = 0 , \quad (7)$$

where the expressions vanish because each state for which $\sigma_r = 1$ has exactly one opposing state for which $\sigma_r = -1$, contributing equally in size. Assuming this stationary point is a minimum we can locate, the pseudo-likelihood approach to inferring an Ising model is exact in the limit of large sample size, and is in this sense qualitatively different from other approximate inverse Ising schemes.

Pseudo-likelihood maximization with l_1 -regularization: Ravikumar, Wainwright and Lafferty in [8] introduced a l_1 -regularized version of the pseudo-likelihood approach, *i.e.* where the functions to be minimized are $[f_r(h'_r, \mathbf{J}'_r) + \lambda \|\mathbf{J}'_r\|_1]$ with some non-zero penalty parameter λ . l_1 (absolute value)-regularization is widely used to recover sparse signals [20–22], in situations where a large fraction of parameters is known to be zero, but not which ones are. The numerical minimization can be done efficiently using convex programming, such as the interior point method of Koh, Kim and Boyd [23], which we have used below. If the goal is to recover the interactions J_{ij} as such, then the l_1 -regularization only introduces a bias and a reconstruction error. If on the other hand the goal is to find which interactions are non-zero, and their sign, and if the interaction graph is known to be sparse, then l_1 -regularization is an important tool. For Ising models without external fields, $\mathbf{h} = \mathbf{0}$, [8] establishes detailed conditions on λ and the scaling parameters (B , N , maximum node degree of the underlying graph d , minimum value of non-zero interactions) for complete such sign-sparsity retrieval with high probability.

Results for high-quality data: Assuming we can find the discussed minimum of (4), the estimator is consistent and the fitting is essentially only limited by imperfect sampling (noisy data). We examine numerically the algorithm's performance for large values of B (using $\lambda = 0$) in the setting of the dilute Sherrington-Kirkpatrick (SK) model [24]. Every J_{ij} is thus non-zero with probability p , and if so drawn from a Gaussian distribution with zero mean and variance $1/c$, $c = pN$. External fields are assumed zero. Reconstruction error is measured by

$$\Delta = \frac{1}{1/\sqrt{N}} \sqrt{\frac{\sum_{i < j} (J_{ij}^* - J_{ij})^2}{N(N-1)/2}} . \quad (8)$$

Without regularization, the minima can be found using, for example, a standard Newton method. Figure 1 shows simulation results for $N = 64$ compared to naive mean-field (nMF) *i.e.* $J_{ij}^{nMF} = -\frac{1}{\beta} (c^{-1})_{ij}$. The curves are the averages of five different runs (error bars are small enough to be omitted). MC sampling (with a basic acceptance/rejection updating rule) was performed us-

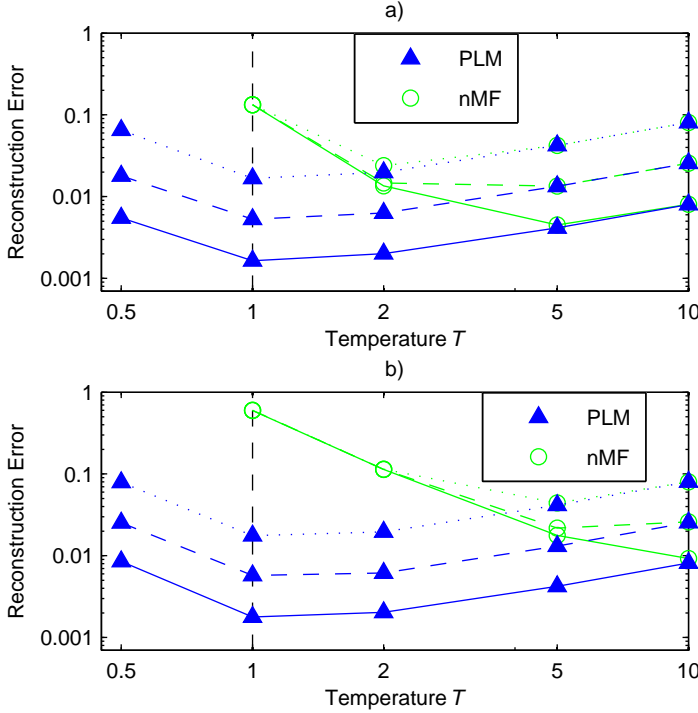


FIG. 1. (Color online). Reconstruction errors of pseudo-likelihood maximization (PLM) and nMF versus temperature for a) fully ($p = 1$) and b) sparsely ($p = 0.1$) connected SK-systems of size $N = 64$. The number of MC samples used are 10^6 (dotted), 10^7 (dashed) and 10^8 (continuous).

ing a warmup time of $10^7 \cdot N$ spin updates and a sampling frequency of one observation every $10 \cdot N$ updates. Evidently, pseudo-likelihood maximization outperforms nMF in the low temperature region. As T approaches one from above (towards the spin glass phase), the naive mean-field method gives poor results, while our logistic regression algorithm appears unaffected. Lowering the temperature further to $T = 0.5$, where nMF and indeed all approximate methods tested on this example to date are unusable, pseudo-likelihood maximization continues to function adequately. On the other hand, as the temperature increases, states become more equiprobable and greater sample sizes are required to extract relevant information about the parameters, resulting in the joining of the curves of the two methods at high T . Performance is at that point limited by the finiteness of B rather than by method choice. Moreover, one can observe that the decrease of Δ seems to follow $\sim \frac{1}{\sqrt{B}}$. Finally, the switch to sparse \mathbf{J} clearly worsens the performance of nMF, but does not seem to affect the pseudo-likelihood scheme. The results for system sizes $N = 16$ and $N = 128$ are similar (data not shown).

Results for low-quality data: Rebuilding the sparsity pattern of \mathbf{J} from few samples using the pseudo-likelihood maximization (PLM) idea has been done nu-

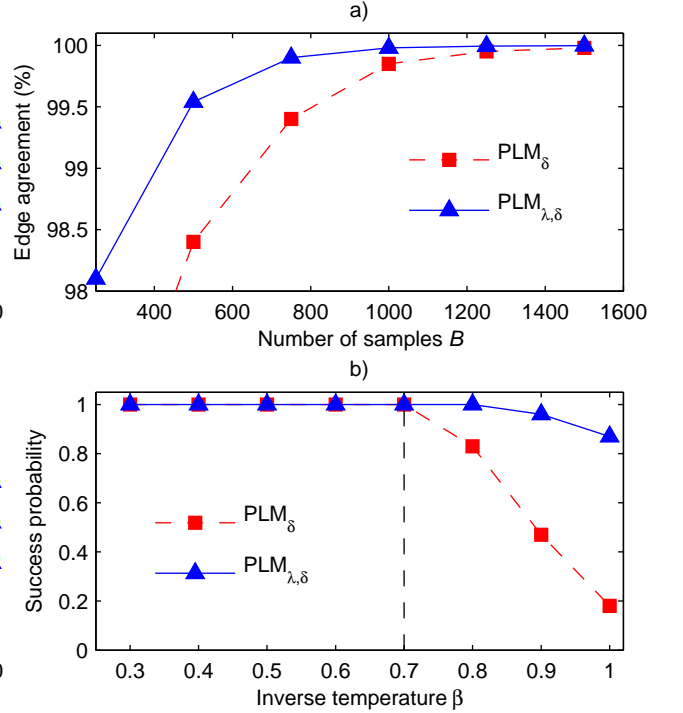


FIG. 2. (Color online). a) Edge agreement versus sample size in a binary SK model of size $N = 100$ and sparsity $p = 0.05$ for PLM_δ and $PLM_{\delta,\lambda}$. $T = 2$ for all data points. b) Probability of 100% edge agreement versus inverse temperature for PLM_δ and $PLM_{\delta,\lambda}$ using $B = 4500$ on 7×7 nearest-neighbor grids ($N = 49$) with 30% dilution.

merically for various sparsity types in [8] and [25]. We provide here some additional results, specifically regarding the advantages of using a regularization term. Taking $\lambda > 0$ after all makes the optimization problem considerably harder computationally. A simpler approach would be to minimize (4) with $\lambda = 0$ and declare all couplings for which $|J_{ij}| < \delta$ to be zero (for some tolerance δ). Intuitively, inclusion of a regularization term should allow for better utilization of sample information than the simpler tolerance approach. As a test case we look at a version of the SK model where the couplings are not Gaussian but binary, $J_{ij} = \pm \frac{1}{\sqrt{pN}}$ (with equal probability). The inference quality is measured as the percentage of pairs (i, j) where the interaction strength is identified correctly as "+", "0" or "-". PLM using tolerance only and PLM using regularization (as well as a tolerance limit) will be referred to as PLM_δ and $PLM_{\delta,\lambda}$ respectively. Figure 2a shows that for $N = 100$, $p = 0.05$ and $T = 2$, $PLM_{\delta,\lambda}$ fits the edges more accurately and gives perfect reconstruction for fewer samples than PLM_δ . Note that in this example guessing $J_{ij}^* = 0$ for all pairs would result in a 95% edge agreement on average. Optimal values of δ and $\{\delta, \lambda\}$ for each B were determined empirically and used on 20 new parameter sets to yield the averages.

For several sparsity structures the performance of PLM

has been shown to drop as the temperature goes below some T_{crit} even if B is quite large [25]. One such example is $B = 4500$ on 7×7 nearest-neighbor grids with positive couplings, where each edge in the grid is removed with probability 0.3 and the remaining couplings are set to one. The "failure" occurs close to the known critical point for the Ising model on such grids [25], $\beta_{crit} \approx 0.7$ [26]. We applied $PLM_{\delta,\lambda}$ and PLM_{δ} to this problem to see whether combined regularization-tolerance can boost performance at low temperatures. Figure 2b shows the outcome, where optimal δ and $\{\delta, \lambda\}$ for each β were again found empirically and probabilities estimated using 200 new grids. A breakdown is indeed seen for PLM_{δ} around $\beta = 0.7$, but the effect on $PLM_{\delta,\lambda}$ is less pronounced if there at all. Perfect edge recovery, using the latter, is had with high probability far into the low-temperature region. The complete data output (not reported) shows that including the tolerance threshold in $PLM_{\delta,\lambda}$ (as opposed to trusting the regularization term alone to force suitable estimates of J_{ij} to zero), becomes necessary at low temperatures. MC samples in this case were generated using a warmup time of $10^7 \cdot N$ spin updates and a sampling frequency of one observation every $2000 \cdot N$ updates.

Discussion: Our results suggest that the pseudo-likelihood approach allows for accurate inference in Ising models even for large strongly coupled systems. The method relies on utilization of complete data sets, implying that the Ising model is considered a model to fit to data as such, and not as a maxentropy model based on means and correlations.

Our results also confirm that including an l_1 -regularization term is helpful in retrieving sign-sparsity from few samples, allowing for complete graph reconstruction even in low temperature regions. A tolerance threshold in combination with l_1 -regularization seems to be necessary to get the best results. It is reasonable

that heavy regularization may run into problems easier than would milder regularization followed by a tolerance limit. We also add that the code employed at $\lambda > 0$ includes estimates of the external fields, thus not utilizing the knowledge that (in our example) $\mathbf{h} = \mathbf{0}$. It is quite possible that even better results can be obtained if the algorithm optimized without fields in the likelihood functions, especially when fitting from few samples.

The time required to solve the N logistic regression problems is not insignificant. For the $N = 64$ cases with 10^8 samples it takes hours on a standard home PC using Newton decent. However, we also applied a quasi-Newton version where the Hessian was approximated using only every 100^{th} or even 1000^{th} sample since the main hurdle is evaluating the Hessian of the objective function, which depends on all 10^8 samples. This version of the procedure located all the minima successfully in a fraction of the time. Also, several algorithms applicable for logistic regression who are typically much faster than Newton decent are available, such as other quasi-Newton and Conjugate Gradient methods. In the few-sample section, computations naturally run much faster and time is not as big an issue. Therefore, in a region where pseudo-likelihood maximization is particularly interesting (small sample size), it is also computationally efficient and competitive.

ACKNOWLEDGEMENTS

E.A. thanks Martin Wainwright, Toshiyuki Tanaka and Michael Hörnqvist for useful discussions. This work was supported by the Academy of Finland as part of its Finland Distinguished Professor program, Project No. 129024/Aurell.

-
- [1] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, *Nature* **440**, 1007 (2006).
 - [2] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, *PNAS* **103**, 19033 (2006).
 - [3] S. Cocco, S. Leibler, and R. Monasson, *PNAS* **106**, 14058 (2009).
 - [4] A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant, *PNAS* **106**, 22124 (2009).
 - [5] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, *PNAS* **106**, 67 (2009).
 - [6] C. Sander and D. Marks, "3d protein structure from sequence alone," (2011), presentation at Physics and Biological Systems, June 14-17, 2011, Orsay, France.
 - [7] M. Weigt, "Integrating statistical-physics inspired inference with molecular dynamics and mutagenesis: From genomic information to protein (complex) structures," (2011), presentation at Physics and Biological Systems, June 14-17, 2011, Orsay, France.
 - [8] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, *Annals of Statistics* **38**, 1287 (2010).
 - [9] G. Darmon, *C.R. Acad. Sci. Paris* **200**, 1265 (1935).
 - [10] E. Pitman and J. Wishart, *Mathematical Proceedings of the Cambridge Philosophical Society* **32**, 567 (1936).
 - [11] B. Koopman, *Transactions of the American Mathematical Society* **39**, 399 (1936).
 - [12] R. Fisher, *Philosophical Transactions of the Royal Society of London* **222**, 309 (1922).
 - [13] Y. Roudi, J. A. Hertz, and E. Aurell, *Front. Comput. Neurosci.* **3** (2009).
 - [14] H. Kappen and F. B. Rodriguez, *Neural Computation* **10**, 1137 (1998).
 - [15] V. Sessak and R. Monasson, *J. Phys. A: Math. Theor.* **42** (2009).
 - [16] M. Mézard and T. Mora, *Journal of Physiology-Paris* **103**, 107 (2009).
 - [17] E. Marinari and V. V. Kerrebroeck, *J. Stat. Mech.*(2010).

- [18] E. Aurell, C. Ollion, and Y. Roudi, European Physical Journal B **77** (2010).
- [19] S. Cocco and R. Monasson, Phys. Rev. Lett. **106**, 090601 (Mar 2011).
- [20] R. Tibshirani, J. Roy. Stat. Soc., Ser. B **58**, 267 (1996).
- [21] D. Donoho and M. Elad, Proc. Natl. Acad. Sci. USA **100**, 2197 (2003).
- [22] M. Gustafsson, M. Hörnquist, J. Lundström, J. Björkegren, and J. Tegnér, Annals of the New York Academy of Sciences **1158**, 265 (2009), ISSN 1749-6632.
- [23] K. Koh, S. Kim, and S. Boyd, J. Mach. Learn. Res. **3**, 1519 (2007).
- [24] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett **35**, 1792 (1975).
- [25] J. Bento and A. Montanari, NIPS **22** (2009).
- [26] D. Zolin, Phys. Rev. B **18**, 2387 (1978).